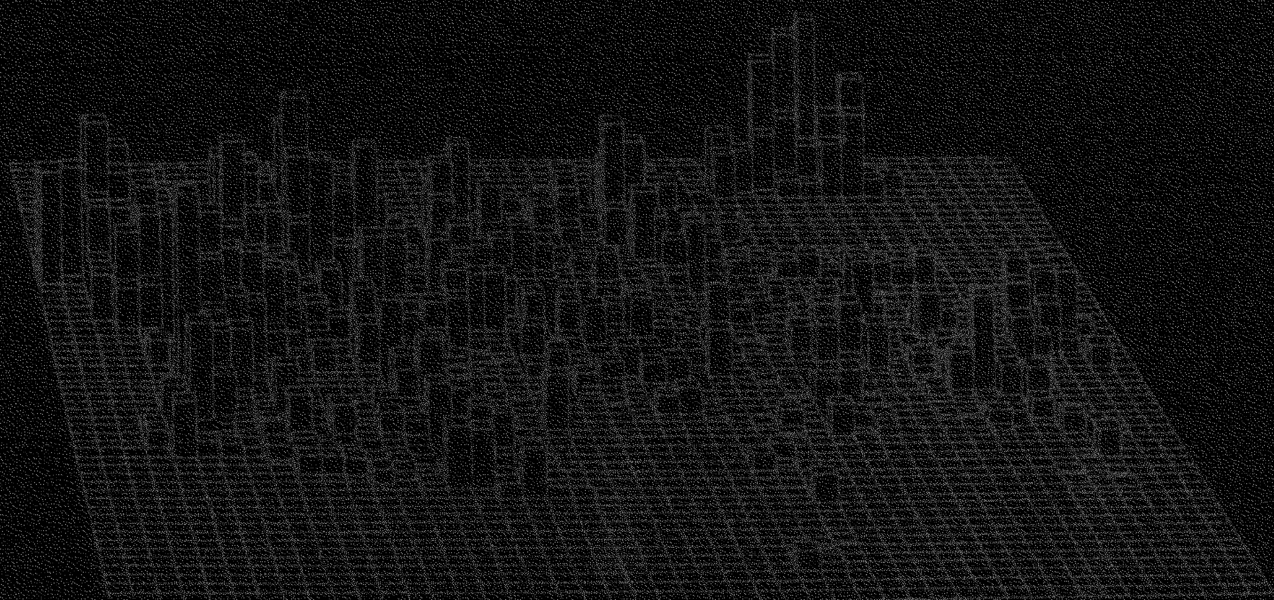
CRU

Chi-square as an alternative to ratios for statistical mapping and analysis

by M. Visvalingam



Census Research Unit
Department of Geography
University of Durham

Working paper 8

The Census Research Unit, Department of Geography, University of Durham, is a small group of research workers investigating aspects of the theory and use of census data. It is currently funded as a research project by the Social Science Research Council.

The diagram on the cover represents total population per 1 km grid square in the northern part of County Durham: the height of each column is proportional to the population in that square. The county is viewed from the west, Gateshead being at the extreme left margin, West Hartlepool at the far right and Bishop Auckland at the centre-right. The original surface was calculated and drawn by computer.

UNIVERSITY OF DURHAM

DEPARTMENT OF GEOGRAPHY

CENSUS RESEARCH UNIT

WORKING PAPER No. 8

NOVEMBER 1976

CHI-SQUARE AS AN ALTERNATIVE TO RATIOS FOR STATISTICAL MAPPING AND ANALYSIS

M VISVALINGAM

CONTENTS

CHI-SQUARE AS AN ALTERNATIVE TO RATIOS FOR MAPPING AND STATISTICAL ANALYSIS

1. INTRODUCTION

Ratios, including such measures as proportions and percentages, are among the most common forms of expression for variables in demography and population geography; they are widely used in both statistical and non-statistical work. This paper suggests that, while preferable to absolute numerical differences in making comparisons between populations of widely varying size, they are still very unsatisfactory numerical measures for mapping and for mathematical and statistical analyses such as regression, correlation and multivariate procedures.

The alternative measure of chi-square (X^2) has several advantages. It takes into account both absolute and relative deviations from the norm, can be represented by a single cartographic symbol and permits a probabilistic approach to the mapping of variables. More important, it demands an assessment of the objectives of a particular investigation before, rather than after, the mechanical stages of the analysis.

Illustrations of the application of the chi-square statistic are, in this paper, confined to a single data set, namely the one-kilometre grid square data on sex composition for County Durham from the 1971 Census. The choice of sex composition was made for a variety of reasons. In the first place, the quantification of sex composition is dependent on only two of the original items of data for each areal unit and thus the effects on the data of the various statistical manipulations under review are more easily discernible. Secondly, the categories of population used are unequivocal, namely the total number of males and the total number of females in each one-kilometre grid square. Thirdly, these are the only items in the Census 100% population record (apart from total population) which are not affected by the arbitrary suppression procedures adopted, for reasons of confidentiality, by the Office of Population Censuses and Surveys. In addition, some pattern of spatial sorting of the sexes could be anticipated (Dewdney and Rhind, 1975, pp. 34-35) and this could be used to test the reliability of the various measures under review.

Most population maps classify and display all the available data. For clarity, however, the illustrations in this paper map only those areas with a marked preponderance of one sex or the other. Although the one-kilometre square data include the entire population, their spatial basis is a sample of all possible incidences of square grids; thus a probabilistic approach can be adopted (Harvey, 1969, p. 284).

Though confined to a consideration of sex composition, the discussion in this paper and the conclusions reached have much wider applications.

2. TRADITIONAL MEASURES OF SEX COMPOSITION

Three measures of sex composition were listed by Shryock and Siegel (1973), namely :

(a) the masculinity proportion, expressed as :

$$\frac{\text{males}}{\text{total population}}$$

(b) the masculinity or sex ratio, expressed as :

$$\frac{\text{males}}{\text{females}}$$

(c) the ratio of the excess or deficit of males to the total population, expressed as :

$$\frac{\text{males - females}}{\text{total population}}$$

These authors were of the opinion that these three measures removed the effects of variations in population size and were therefore suitable for inter-area, inter-group and temporal comparisons. Since the simple numerical excess or deficit of males is affected by the size of the population, Shryock and Siegel considered such a measure inadequate for making comparisons between populations. In addition, these authors noted that the sex ratio, the demographer's principal measure of sex composition, was a "more sensitive" indicator and postulated that this was due to its relatively smaller base.

Since all the three measures of sex composition listed by Shryock and Siegel are derived from the same items of data and are functions of each other, they should give similar descriptions.

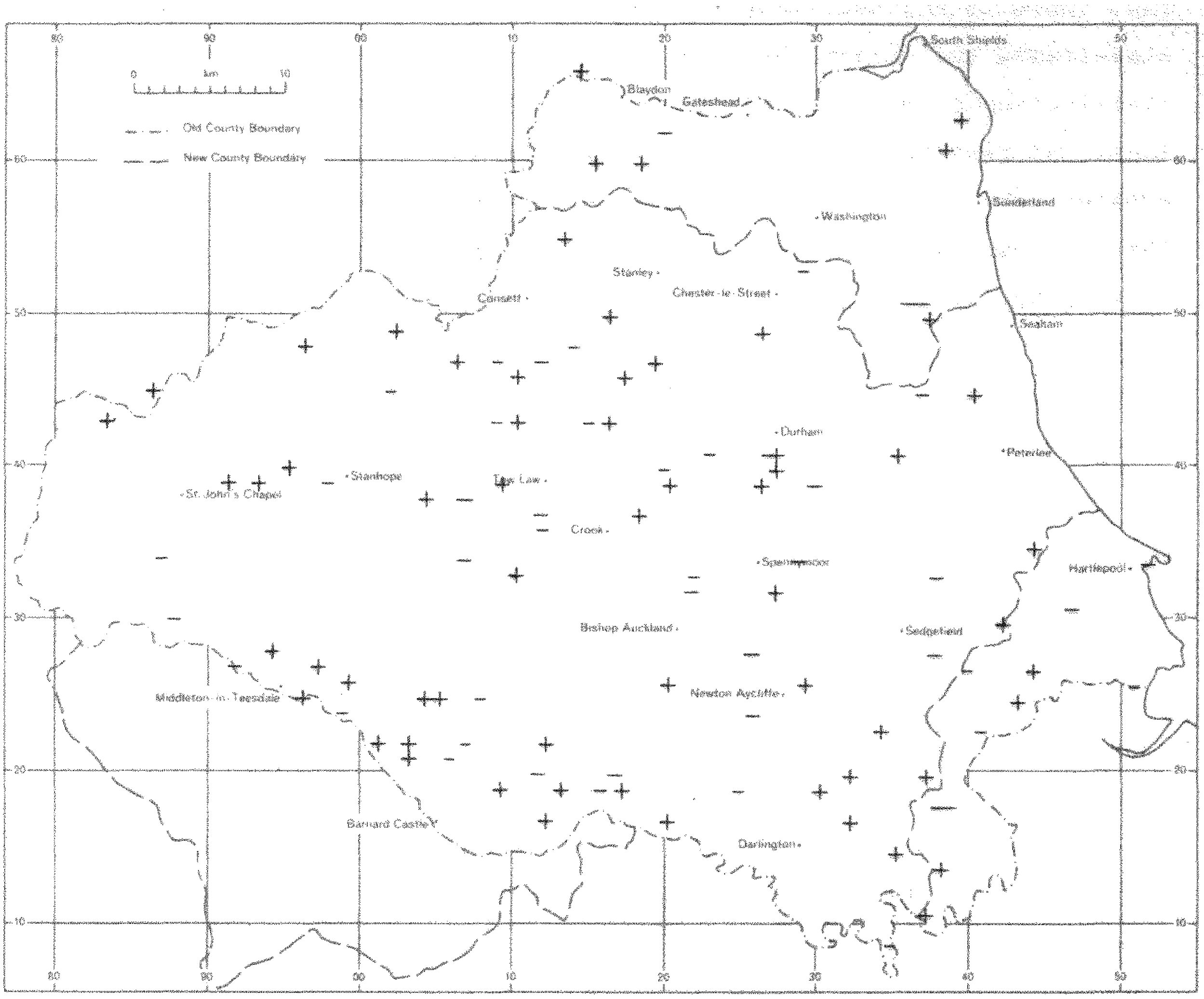
Measure (c) is derived from (a) by :

$$c = a-(1-a) = 2a-1$$

Thus there is a linear relationship between the two. If, instead of classifying the entire range of values, we consider only the spatial distribution of extreme values, both measures (a) and (c) produce identical maps (Fig 1), with 62 squares having masculinity proportions greater than or equal to 0.716 (c=0.432) and 46 squares with values less than or equal to 0.307 (c=-0.398). Of these 108 extreme cases, however, only 13 appear to merit any further investigation (Table 1). The majority of the very high and very low masculinity proportions were the product of very small populations, including several cases of the situations shown in Table 2.

Chance variations in the numbers of males and females in squares with very small populations can lead to very high or very low ratios (Choynowski, 1959). Dewdney and Rhind (1975) have already shown that the effect of population size is not removed by the use of the masculinity proportion, still less by the use of the masculinity ratio. Evans (1976) demonstrated this point and showed the effects of the adjustment procedures employed by OPCS on ratios derived from small populations. He also pointed out some of the implications of using ratios derived from small samples such as those provided by the 10% census data.

Ratios have traditionally been used for the study of macro-populations and for making international comparisons. When applied to micro-populations with a broad size range (as is the case with the one-kilometre grid square data), they have an adverse effect on the ranking of those populations by producing distributions whose extremes are dominated by squares with small populations. This has serious implications, not only for mapping but also for multivariate statistics which draw on correlation structures. From Figure 1 it is obvious that measures (a) and (c) are strongly biased towards the discrimination of sex differences in rural areas; it is apparent that, while exaggerating small differences in such areas, these measures conceal differences in urban areas which may be much more important because of the large numbers of people involved.



	% males	1 km squares
+	>71·61	62
	30·07–71·61	1948
–	<30·07	46

Figure 1. Distribution map of extremes of masculinity proportion in County Durham

TABLE 1 EXTREME MASCULINITY PROPORTIONS WHICH APPEARED TO MERIT FURTHER INVESTIGATION

Kilometre square	Number of males	females	a	diff.	X_s^2
4380 5600	52	18	0.743	34	16.51
4260 5480	67	13	0.838	54	36.45
*4270 5480	20	8	0.714	12	5.14
4110 5460	2	10	0.167	-8	-5.33
*4270 5420	2458	1020	0.707	1438	594.55
4260 5400	95	229	0.293	-134	-55.42
4270 5400	227	50	0.819	177	113.10
3970 5380	7	20	0.259	-13	-6.26
4260 5380	61	16	0.792	45	26.30
4060 5370	17	58	0.227	-41	-22.41
*4170 5340	21	44	0.323	-23	-8.14
4210 5320	10	52	0.161	-42	-28.45
4290 5250	126	7	0.947	119	106.47
*4230 5230	85	38	0.691	47	17.96
4160 5210	5	12	0.294	7	-2.88
*4070 5180	132	57	0.698	75	29.76
4090 5180	273	8	0.972	265	249.91
4150 5180	7	17	0.292	-10	-4.17
*4060 5150	10	21	0.323	-11	-3.90
*4080 5150	19	8	0.704	11	4.48
*4310 5110	10	23	0.303	-13	-5.12

*Squares not represented in Figure 1

TABLE 2 EXAMPLES OF EXTREME MASCULINITY PROPORTIONS DUE TO SMALL POPULATION SIZE

Total	Males	Females	Masculinity
3	3	0	1.000
6	5	1	0.833
5	4	1	0.800
4	3	1	0.750
7	5	2	0.714
10	3	7	0.300
7	2	5	0.286
5	1	4	0.200
2	0	2	0.000

Measures (b) and (a) are related to each other by the following relationships :

$$a = \frac{b}{1+b} \qquad \text{and} \qquad b = \frac{a}{1-a}$$

Figure 2 illustrates these relationships, using the Durham data set, and shows that the effect of population size must be similar in both cases. As (b) is bounded by zero and is asymptotic towards the vertical line a=1, it produces a very skewed distribution of values, progressively exaggerating the arithmetic differences above and compressing those below the value of 1. Thus, while the sex ratio (b), when used with an arithmetic scale, identifies the predominance of males more sharply, the inverse (females/males) would be more useful for the study of female preponderance. While either form of the sex ratio is quite adequate for the comparison of national figures, which are generally very close to the norm, neither is adequate (without transformation) for use over a broader range of population size. Moreover, like the other measures, the sex ratio is not a suitable input to multivariate statistical analysis as the extremes, which are largely the products of small populations, are bound to dominate the calculation of correlation coefficients.

Figure 3 and Table 3 demonstrate that the use of consistent ratio cut-offs is likely to mask quite large absolute differences in areas with large populations. As Table 3 shows, the numerical differences between the sexes required to produce a particular masculinity proportion vary directly with population size.

TABLE 3 NUMERICAL DIFFERENCES BETWEEN THE SEXES REQUIRED TO PRODUCE A MASCULINITY PROPORTION OF 60% AT DIFFERENT POPULATION SIZES

Total Population	Males	Numerical difference between sexes
10	6	2
100	60	20
1,000	600	200
10,000	6,000	2,000
100,000	60,000	20,000

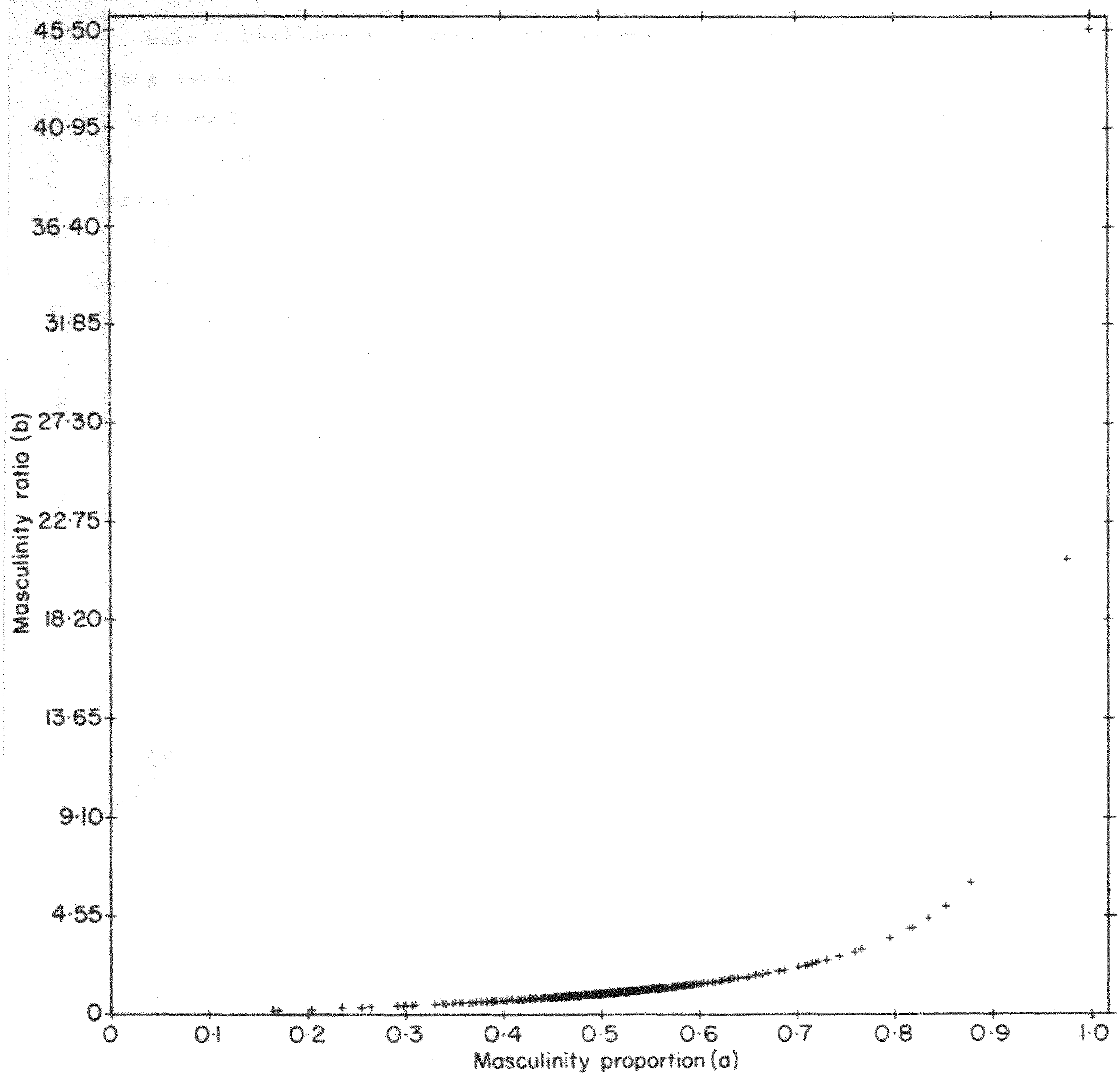


Figure 2. Relationship between masculinity proportion and sex ratio

Thus, although the scatter (variance) increases with population size (Fig. 3), areas with large numerical differences between the sexes are not likely to be picked out using a constant sex ratio away from the origin, since this assumes that deviation is a linear function of population size. Even if the tolerance (or cut-offs) in terms of ratios were to be derived as an inverse function of population size, ratios would still be unsuitable for comparing populations of greatly differing sizes. Equally, the use of absolute numerical differences alone is inadequate, since it is likely to be strongly biased towards the discrimination of sex differences in areas with large populations (in most cases urban areas), while ignoring quite large relative differences in small populations. Analyses based on the paired-t test showed that a greater proportional difference (in relation to population size) is to be expected in smaller populations than in larger ones.

3. AN ALTERNATIVE MEASURE ?

What, then, are the alternatives ? Dewdney and Rhind (1975) removed the worst anomalies by considering only those one-kilometre squares containing more than 10 people, but the discussion so far makes it clear that a distorted ranking and distribution still exists. If possible, suppression cut-offs should be selected on a more objective basis. One problem apparent from Figure 3 is that the bivariate relationship between males and females lacks homoscedasticity and thus the various t-tests are not applicable. To some extent this can be circumvented by considering subsets of large, small and possibly also middle-sized populations separately so that expectations of variance can be more legitimately standardised for the range of population size in each subset. However, not only is the delimination of subsets likely to be subjective, if not arbitrary, but the boundaries themselves will also vary with different ranges of populations and groups, making comparative studies more difficult. For example, to cope with problems arising from heteroscedasticity, different numbers and/or thresholds may be necessary for the delimitation of "small" and "large" populations in different data sets (those of County Durham and Great Britain respectively, for example). Thus, if at all possible, it is desirable that the selected measure should discriminate over the entire range of values by making suitable but consistent internal adjustments.

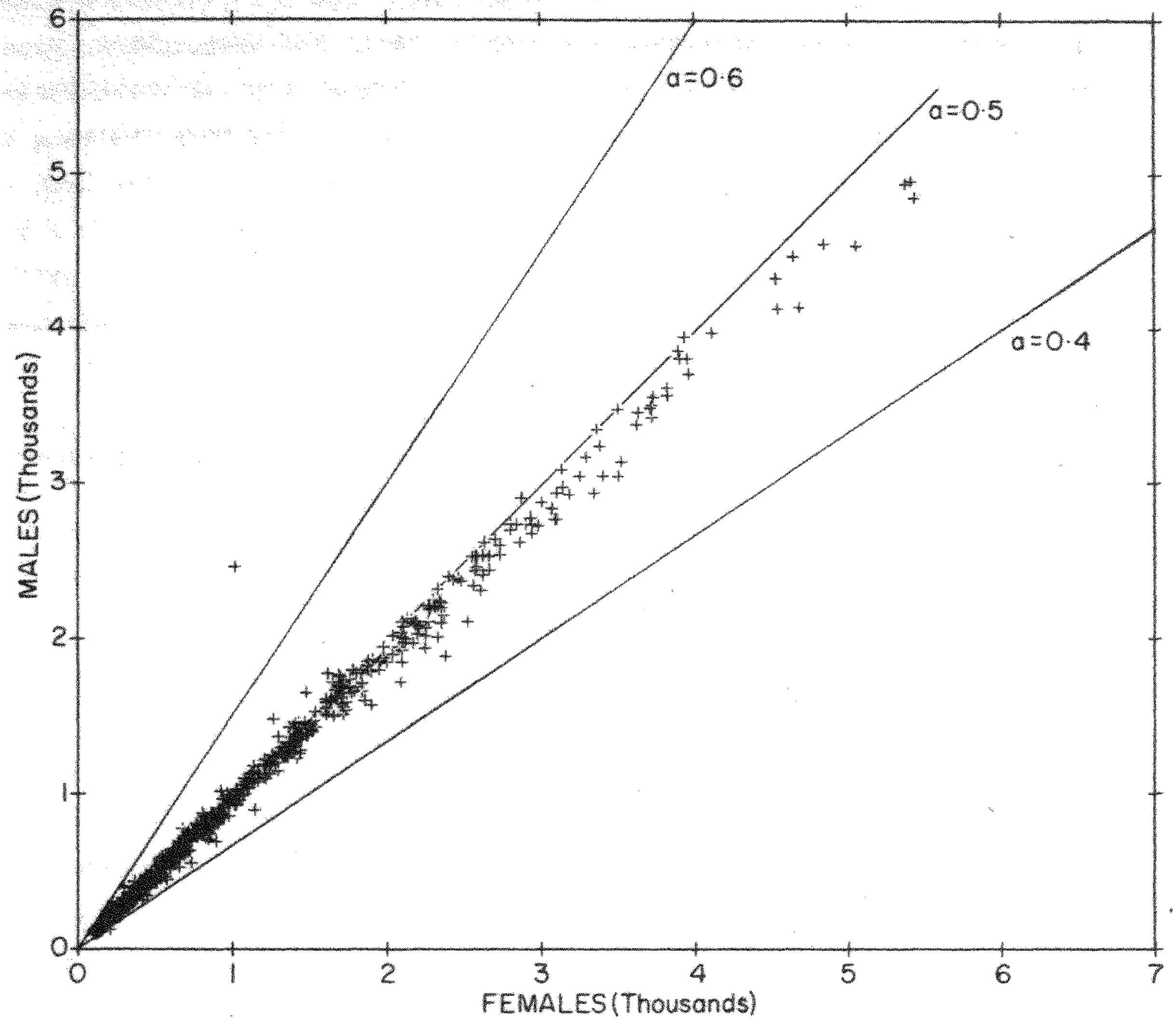


Figure 3. Numerical differences expected in a range of population sizes by constant masculinity proportions (a); County Durham data with suppression of all one-kilometre squares containing less than 100 males or females

The chi-square statistic was investigated as an alternative measure of sex composition which might overcome many of these problems. Here the approach is one of projecting an "expected" frequency and calculating the magnitude of the (absolute x relative) departure of observed values from that expectation :

$$X^2 = \sum \left[(O-E) \times \left(\frac{O-E}{E} \right) \right] = \sum \left[\frac{(O-E)^2}{E} \right]$$

where O = observed frequency

E = expected frequency

The significance of this departure can then be assessed at different levels of confidence. If we initially assume a norm of numerical equality between the sexes, then E = half the observed total population, and the expected value (E) for a particular area with observed males O_m and observed females O_f can easily be derived as $(O_m + O_f)/2$. The departure of O_f, O_m would have to include not only the departure of O_m from expectation (giving D_m) but also that of O_f (giving D_f). Thus X^2 for each area includes two terms :

$$X^2 = \frac{(O_m - E_m)^2}{E_m} + \frac{(O_f - E_f)^2}{E_f}$$

In this particular case, $E_m = E_f$ and thus $D_m^2 = D_f^2$. Thus :

$$X^2 = 2\frac{(O_m - E_m)^2}{E_m} = 2\frac{D_m^2}{E_m}$$

For a given masculinity proportion, X^2 values are much higher for larger populations than for smaller ones (Table 4).

TABLE 4 X^2 VALUES FOR A MASCULINITY PROPORTION OF 60% AT DIFFERENT POPULATION SIZES

Total	E_m	O_m	D_m^2	X^2
10	5	6	1	0.4
100	50	60	100	4.0
1,000	500	600	10,000	40.0
10,000	5,000	6,000	1,000,000	400.0

As a corollary, the same absolute difference is rated as more important, both conceptually and numerically, in a smaller population, since the denominator (E_m) is smaller. For example, while a difference of 10 in a population of 10,000 gives a X^2 value of 0.04, the same difference in a population of 100 (refer second row of Table 4) results in a X^2 value of 4.0. Thus, although the chi-square statistic expects larger numerical differences in larger populations, unlike ratios it does not expect the magnitude of difference to be a linear function of population size.

A better understanding of the chi-square measure may be obtained if the chi-square formula for equal probabilities of males and females is inverted to derive the cut-off in terms of an observed value O_m for a given population. (Note that this inversion is not universally applicable). In this case :

$$O_m = E_m \pm \text{tolerance} = E_m \pm \sqrt{\frac{X^2}{2} \cdot E_m}$$

Hence, for a given X^2, the variation that is permitted is basically a square root function (see Fig 4) of the expected number of males (i.e. of population size). In practice, a value of X^2 is chosen to define the tolerance numerically, while the root function keeps the variation at an acceptable level. With one degree of freedom, X^2 tables indicate values of 3.84 and 6.63 at the 95 and 99 per cent confidence levels respectively and crude representations of the corresponding cut-offs are shown in Figure 5 (cf. Fig. 3).

This measure automatically places less importance on very small populations. With 3 people in a one-kilometre square, even if they are all males, X^2 (for simplicity ignoring Yates' correction) is not significant at the 95 per cent confidence level because :

$$X^2 = \frac{2(3-1.5)^2}{1.5} = 3.0$$

However, it would be incorrect to say that X^2 values can be interpreted without reference to population size. The effect of chance fluctuations (for example, due to the positioning of the grid) is greater with smaller populations than with larger ones. Moreover, in general, X^2 values are unreliable when values for expectation are less than 5.

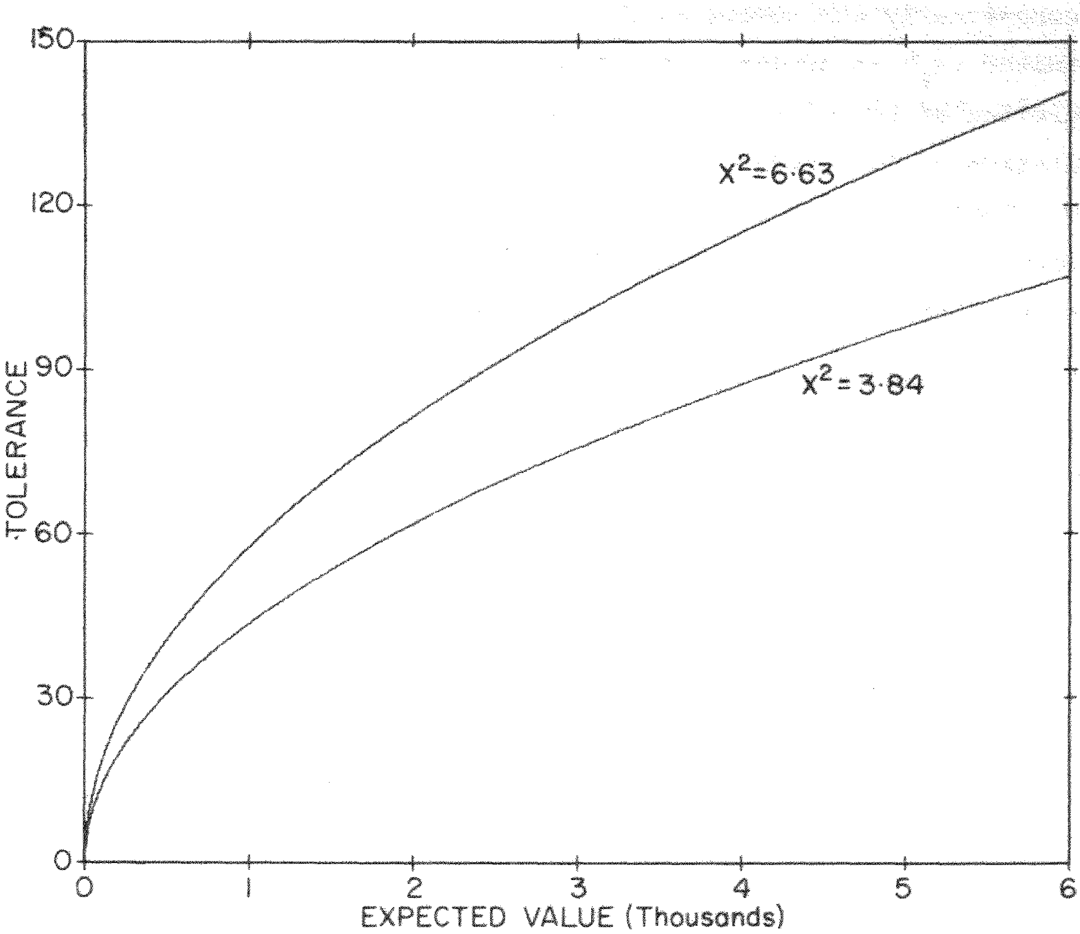


Figure 4. Tolerance as a function of expectation for given values of X^2

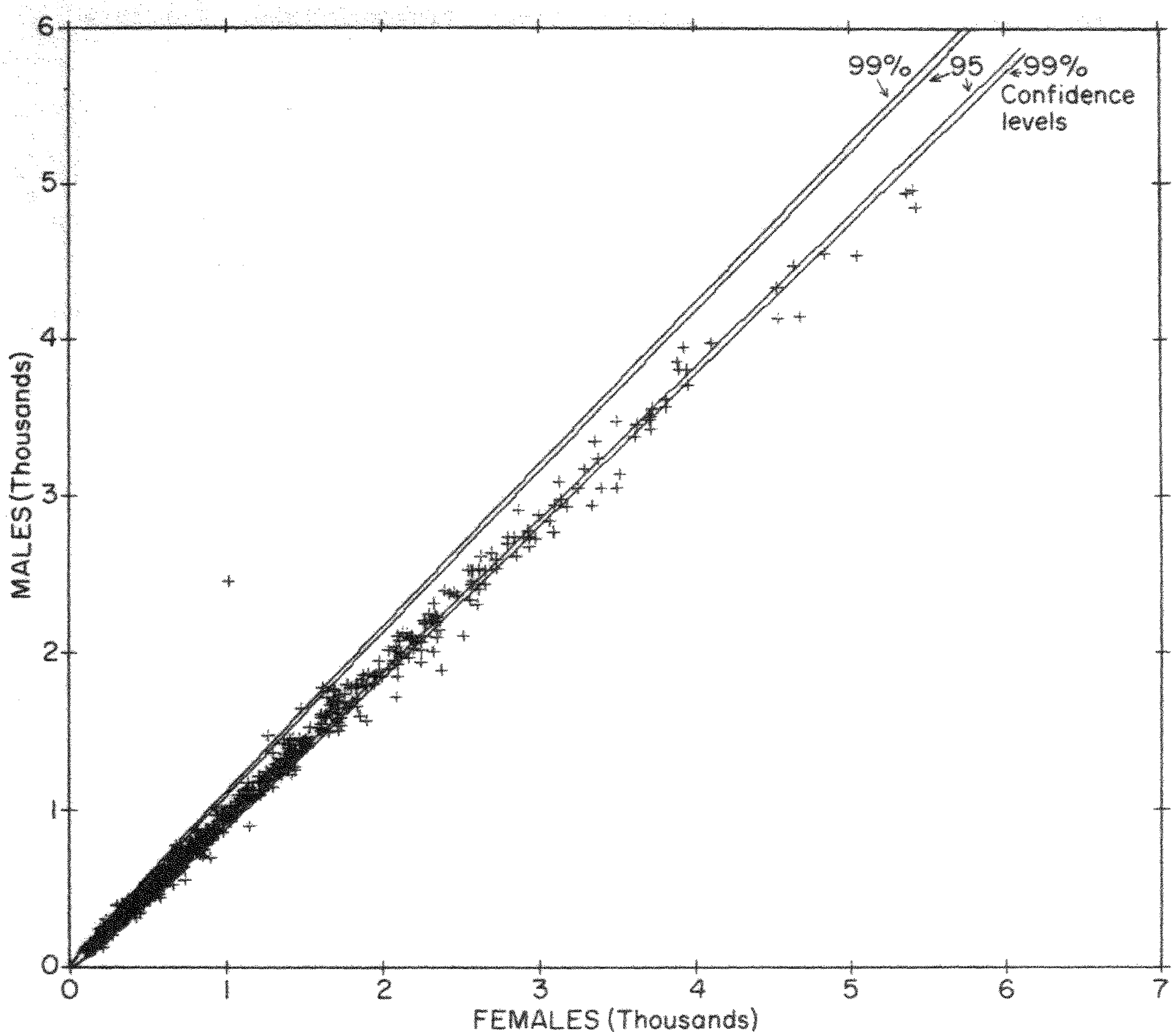


Figure 5. 99 and 95 per cent confidence limits of X^2 for one degree of freedom

In the study of sex composition, it is valuable to distinguish between areas with male as opposed to female preponderance. For this purpose, X^2 may be modified as follows:

$$X_s^2 = \operatorname{sgn}(O_m - E_m).\ X^2$$

where the sign function sgn ( ) is the sign, + or -, of the expression in parentheses. Thus male and female preponderance are indicated by positive and negative values of X_s^2 respectively; the numerical values remain exactly the same as for the standard X^2 statistic. Figures 1 and 6 show that the spatial pattern of extremes in the masculinity proportion and the X_s^2 values are unrelated (see also Table 1). Although at the 95 per cent confidence level there is a higher probability of errors in judgement, the additional cases confirm and extend the pattern reported at the 99 per cent confidence level. As, in 1971, County Durham had only 721,558 males as against 753,818 females (i.e. a difference of 32,360 and a masculinity proportion of 0.489) one would expect a greater number of one-kilometre squares with female than with male preponderance (Table 5), given an expectation of equal numbers of males and females.

TABLE 5 NUMBERS OF GRID SQUARES IN COUNTY DURHAM WITH MALE OR FEMALE PREPONDERANCE AS ASCERTAINED BY THE X_s^2 METHOD

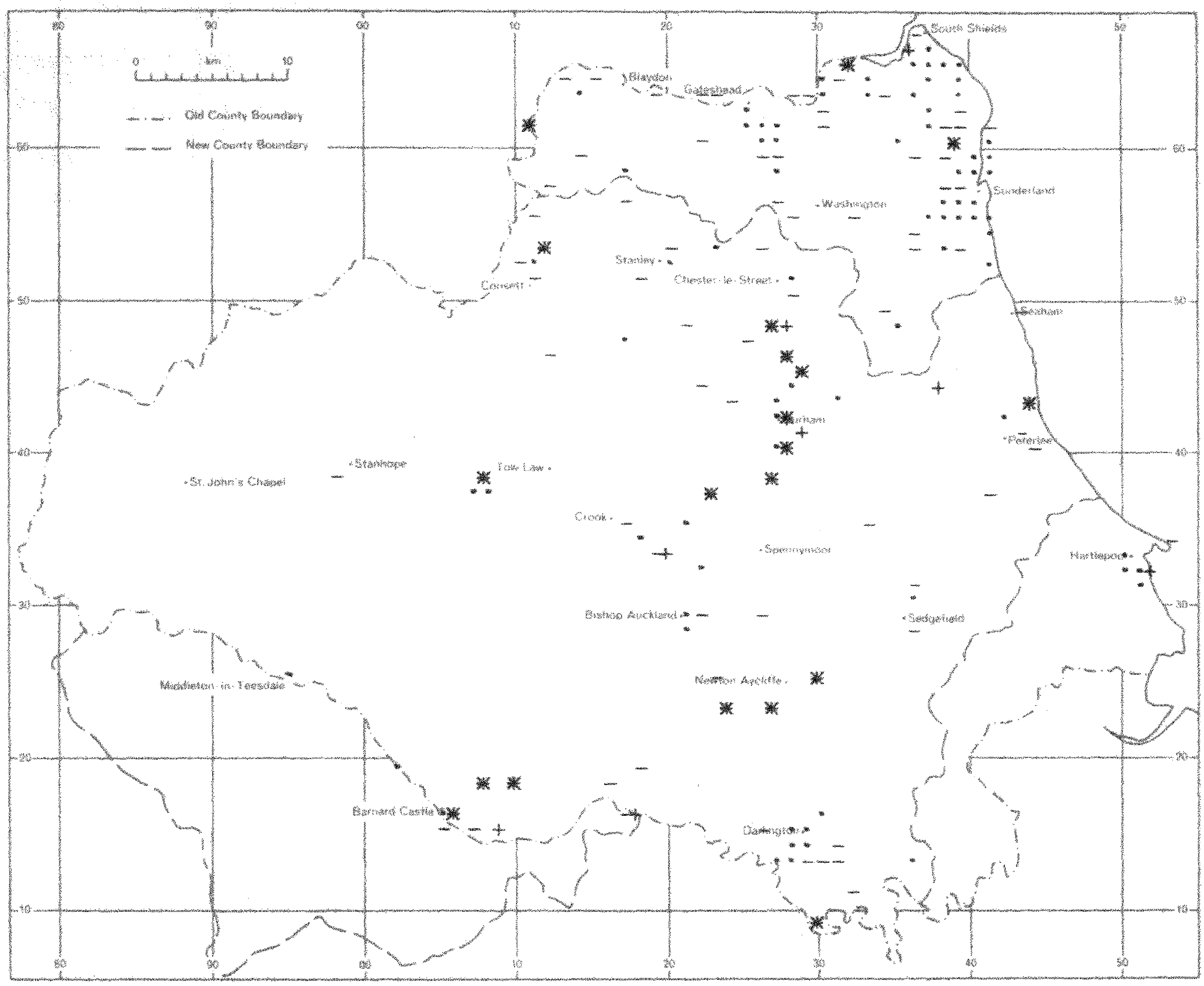
Expectation	95% confidence level		99% confidence level	
(% of total)	Female	Male	Female	Male
50.0	149	27	79	21
48.9	54	44	30	26

If the expectations for males and females were based on the County proportions :

$$X^2 = \frac{(O_m - 0.489T)^2}{0.489T} + \frac{(O_f - (1-0.489)T)^2}{(1-0.489)T}$$

where T = the total number of people in the one-kilometre grid square.

Figure 7 shows the areas of male and female preponderance at the 99 and 95 per cent confidence levels. The new expectation of 48.9 per cent, compared with that of 50 per cent, indicates more one-kilometre squares with a significant proportion and number of males (Table 5).



* Male preponderance at 99% significant level

\+ Male preponderance at 95% significant level

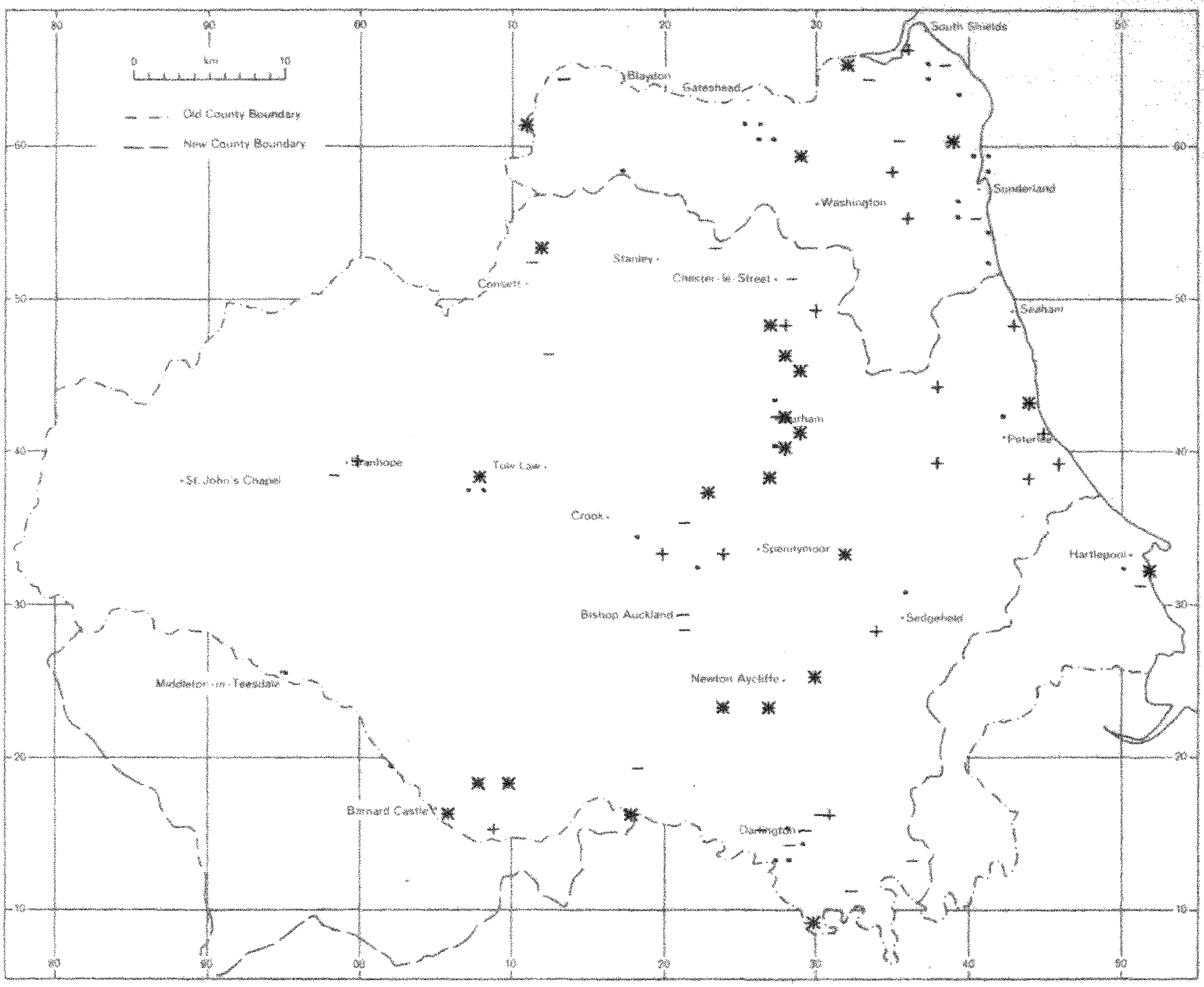
Insignificant deviation from 50% males

\- Female preponderance qt 95% significant level

. Female preponderance at 99% significant level

(Where symbols are printed outside the boundaries of County Durham, e.g. just north of Sunderland, the squares to which these symbols refer lie partly within the County)

Figure 6. Distribution map of sex composition in County Durham based on X_s^2 values and an expectation of 50 per cent males



* Male preponderance at 99% significant level

+ Male preponderance at 95% significant level

Insignificant deviation from 48·9% males

– Female preponderance at 95% significant level

. Female preponderance at 99% significant level

(Where symbols are printed outside the boundaries County Durham, e.g. just north of Sunderland, the squares to which these symbols refer lie partly within the County)

Figure 7. Distribution map of sex composition in County Durham based on X_s^2 values and an expectation of 48.9 per cent males

There are also fewer squares with a preponderance of females at the new level of expectation.

Even with expectations based on the County average, there are more squares with a preponderance of females than of males, suggesting that the male concentrations are more extreme and less frequent than the female. Local disparities between Figures 6 and 7 are observable, notably between Peterlee and Seaham, around Spennymoor and in the north-east of the County. More detailed analyses of inter-group variations in sex composition would be necessary to verify conjectures concerning causal factors and processes, and it would be interesting to see if and how the spatial pattern of sex composition changes with more refined contingency tables involving other variables from the 100% and 10% census data.

4. DISCUSSION

4.1. Formulation of Expectation

"Significance may be measured only in terms of some *a priori* expectation" (White,1972). The key issue in the use of chi-square is thus the formulation of expectation. Values for expectation are usually derived mechanically from contingency tables or related to some probability distribution such as the Binomial, Poisson (Choynowski, 1959; White, 1972) or Normal (Norcliffe, 1972) distributions (Lieberman,1971). In the study of sex composition, a data-independent and arbitrary norm of 50 per cent males is equally valid, but it assesses the significance of deviations with respect to a different null hypothesis. A change in expectation from 50 per cent males to the County average of 48.9 per cent did not make a great difference to the ranking of one-kilometre squares in County Durham. However, when the variable under consideration has a skewed distribution, the degree of deviation and the ranks of the kilometre squares alter markedly with a change of expectation. Thus the application of chi-square to variables such as unemployment and overcrowding introduces complications, since expectations can also, in principle,be prescribed on the basis of social theories and political objectives; such expectations would, however, be difficult, if not impossible, to define numerically. It is imperative, in such cases, that the reasons for selecting a non-statistical norm are clearly stated in order that results are not misinterpreted.

4.2. Mapping of Chi-square Values

While probability mapping proved convenient for the identification of areas with extreme male or female preponderance, it might be more appropriate in other instances to classify and map X_s^2 values for all data locations. This is likely to be more useful when expectations are derived on the basis of non-statistical criteria, resulting in a skewed distribution of X_s^2 values.

4.3. Unreliable Values for Small Populations

With one degree of freedom, the chi-square measure is considered unreliable when the expected number of occurrences in any cell of the contingency table is less than 5 (Siegel, 1956, p. 46). The conventional solution is to aggregate data for adjacent locations and to ascertain whether a similar spatial pattern results. As has already been shown, ratios are also most unstable for small populations. The spatial pattern of sex composition at different levels of aggregation is currently being investigated (using a subset of the 1971 one-kilometre grid-based census data for Great Britain) by the present author and J.C.Dewdney, using both chi-square and ratio measures. The results will be reported in a later Working Paper.

4.4. Events in the Tails of a Distribution

A problem general to all forms of probability mapping is that of separating non-random from random events in the tails of a distribution (Norcliffe, 1972; Visvalingam, 1972). It is hoped that the additive property of chi-square can be utilised as another means of assessing the significance of variation in the spatial domain.

4.5. Variables for Multivariate Analysis

Variables expressed in ratio form are likely to distort multivariate analysis of micro-populations because the extreme values, which are largely the product of small populations, would dominate the calculation of correlation co-efficients. A comparison of X_s^2 values with ratio correlations (with and without population weighting) is to be undertaken by I.S. Evans and the present author in an attempt to derive the most meaningful expressions for use in multivariate statistics. Results will be reported in a later Working Paper.

4.6. Open Ratios

Chi-square would appear to be a satisfactory alternative to the traditional measures of sex composition. This is partly because the total population can be simply categorised in terms of those who are males and those who are not (i.e. females). This is an essential pre-requisite for the use of the chi-square measure (Lieberman, 1971, section V, p. 301). Variables expressed as closed ratios, giving values from zero to one inclusive (for example, retired persons as a proportion of total population, or those seeking work as a proportion of the economically active population) can be re-expressed to take the X_s^2 form. Although some open ratios can legitimately be re-defined in a closed ratio form (for example, the masculinity proportion can be derived from the masculinity ratio), other open-ratio variables, such as persons per room, room per household or the age index, need further consideration.

4.7. Other Merits of the Chi-square Model

Chi-square can be applied to spatially discrete phenomena without reference to conditions of normality or homoscedasticity and it should apply over the entire range of values, becoming suspect only when the expected frequency falls below 5. Unlike parametric tests (for example, the use of the z and t statistics) for mapping, which base estimates of error on the standard deviation, tolerance is evaluated by chi-square as a function of the central tendency of the data set, which is less influenced by errors in the data. Abnormally high values of chi-square would have to be checked but, in general, erroneous data at specific locations would not directly influence the rating of adjacent locations, except when chi-square values were pooled (see 4.4 above).

5. CONCLUSION

Ratios and differences are known to be inadequate measures for the study of micro-populations. Chi-square appears to be an adequate alternative for the study of variables which can be expressed as closed proportions. The X_s^2 value is a directional measure of both absolute and relative deviations from expectation and lends itself to significance testing and probability mapping. Comparative studies undertaken using the Great Britain data set confirm the arguments put forward in this paper. At the same time, more extensive studies are being conducted into the applicability of chi-square to other variables, such as unemployment, retirement and overcrowding, and its suitability for multivariate analysis.

ACKNOWLEDGEMENTS

The author is grateful to other members of the Census Research Unit for their comments on the paper, to Ray Hudson for discussions on the selection of non-statistical norms, and especially to Ian Evans and Nick Cox for many stimulating arguments and helpful suggestions during the drafting process. Any remaining errors are, of course, the author's responsibility.

In addition, thanks are due to Mrs. Joan Dresser who typed the manuscript, Derek Hudspeth and Arthur Corner, who prepared the diagrams for publication, John Normile for the reprographics and John Dewdney for his editorial work on the manuscript.

REFERENCES

CHOYNOWSKI, M. (1959) "Maps based on probabilities", Journ. Am. Statist. Ass., 54, pp. 385-8

DEWDNEY, J.C. and RHIND, D.W., eds, (1975) People in Durham - a census atlas, Census Research Unit, Dept. of Geography, University of Durham, pp. 34-5

EVANS, I.S. (1976) "Percentages, ratios and inbuilt relationships", Statistics Notes to Durham 1H Students

HARVEY, D. (1969) Explanation in Geography, Edward Arnold, London, p. 284

LIEBERMAN, B. (1971) Contemporary Problems in Statistics, Oxford University Press, Section V, pp. 279-346

NORCLIFFE, G.B. (1972) "Probability mapping of growth processes", Economic Geography, 48, pp. 428-438

SHRYOCK, H.S., SIEGEL, J.S. and Associates (1973) Methods and Materials of Demography, U.S. Govt. Printing Office, Washington, D.C., Vol.1, pp. 188-200

SIEGEL, S. (1956) Non-parametric Statistics for the Behavioral Sciences, McGraw-Hill, Tokyo, p. 46

VISVALINGAM, M. (1972) "An application of statistical hydrology to an evaluation of some methods of measuring the water table", Unpub. Ph.D. Thesis, University of Hull, pp 146-148

WHITE, R.R. (1972) "Probability maps of leukaemia mortalities in England and Wales" in Medical Geography, Techniques and Field Studies (ed. N.D.McGlashan) Methuen, London, pp. 173-185